Asymptotic properties of mixture-of-experts models

Madalina Olteanu, Joseph Rynkiewicz

Asymptotic properties of mixture-of-experts models

Olteanu, M., Rynkiewicz, J.,
SAMM, EA 4543, Universite Paris 1,
90 Rue de Tolbiac, 75013 Paris, France

December 16, 2010

Abstract

The statistical properties of the likelihood ratio test statistic (LRTS) for mixture-of-expert models are addressed in this paper. This question is essential when estimating the number of experts in the model. Our purpose is to extend the existing results for simple mixture models (Liu and Shao, [13]) and mixtures of multilayer perceptrons (Olteanu and Rynkiewicz, [14]). In this paper we first study a simple example which embodies all the difficulties arising in such models. We find that in the most general case the LRTS diverges but, with additional assumptions, the behavior of such models can be totally explicated.

keyword Mixture of experts, likelihood ratio statistic test, asymptotic statistic

1 Introduction

Derived from neural networks literature, Mixtures of Experts (ME) (Jacobs et. al., [9]) and Hierarchical Mixtures of Experts (HME) (Jordan and Jacobs,[11]) generalize linear regression models. HME are mixtures of "experts" (for example, linear regression models) organized in a tree-structured network. The network assigns a weight to each expert and then produces an output which combines the outputs produced by all experts according to

their weights. Unlike mixtures of regression models, the weights depend on the input x. The ME discussed in this paper is a particular case of HME, where the network has only one layer. The conditional density of a ME can be generally written as:

$$g\left(y|x,\phi\right)=\sum_{i=1}^{p}\pi_{\nu_i}(x)g_{\theta_i}(y|x),$$

where $\phi=\left(\nu_1^T,...,\nu_p^T,\theta_1^T,...,\theta_p^T\right)$ is the parameter of the model. Usually, the weights or "gating functions" are chosen to be logistic type

$$\pi_{\nu_i}(x)=\frac{\exp\left(\nu_i^T x\right)}{\sum_{j=1}^{p}\exp\left(\nu_j^T x\right)},$$

while g_θ may be Poisson, Binomial or Gaussian distributions.

When the model is assumed to be correctly specified, the maximum likelihood estimates converge to the true values of the parameters and are normally distributed (Jiang and Tanner,[10]). However, the true model is not usually known and the true parameter is unidentifiable. This paper studies the asymptotic behavior of the likelihood ratio test statistic (LRTS) for mixtures of experts and extends the results for simple mixtures models (Liu and Shao [13]). In Section 2, we present the model and illustrate, on a simple example, the cases of divergence and, under some additional assumptions, convergence. Section 3 deals with the asymptotic of the LRTS in a more general frame.

2 The model and a simple example

Let $\left(X_k,Y_k\right)_{k\in\mathbb{Z}}$ be a sequence of independent and identically distributed random vectors defined on a probability space $(\Omega,\mathcal{K},\mathbb{P})$ and let us denote by μ the probability distribution of the vector (X_k,Y_k). Let $\mathcal{P}=\{g_\theta,\theta\in\Theta\}$ be a set of densities with respect to some positive measure λ, where Θ is a finite-dimensional set. Let us consider an observed sample $\{(x_1,y_1),...,(x_n,y_n)\}$ of the sequence (X_k,Y_k). For every x_k, the true density of Y_k conditionally to $X_k=x_k$ is

$$g^0\left(y_k\mid x_k\right)=\sum_{i=1}^{p_0}\pi_{\nu_i{}^0}(x_k)g_{\theta_i{}^0}\left(y_k\mid x_k\right),$$

where $g_{\theta_i^0}\in\mathcal{P}$, $\pi_{\nu_i{}^0}(x_k)\geq 0$, $\sum_{i=1}^{p_0}\pi_{\nu_i{}^0}(x_k)=1$ and $\phi_0=\left(\theta_1^0,...,\theta_{p_0}^0,\nu_1^0,...,\nu_{p_0}^0\right)^T$ is the global parameter of the model. Let us remark that this model is the

general parameterization of mixtures of experts.
The set of possible conditional densities is

$$\mathcal{G} = \Bigg\{ g\left(y_k \mid x_k\right) = \sum_{i=1}^{p} \pi_{\nu_i}(x_k) g_{\theta_i}\left(y_k \mid x_k\right) \ , \ \pi_{\nu_i}(x_k) \geq 0 \ ,$$

$$\sum_{i=1}^{p} \pi_{\nu_i}(x_k) = 1 \ , \ g_{\theta_i} \in \mathcal{P} \ , \ p_0 \leq p \leq P \Bigg\}$$

where $P \in \mathbb{N}^\star$ is fixed, sufficiently large.

For $g \in \mathcal{G}$, let

$$l_n\left(g\right) = \textstyle\sum_{k=1}^{n} \ln \ g\left(Y_k \mid X_k\right)$$

be the conditional log-likelihood function of $((X_1, Y_1), \cdots, (X_n, Y_n))$. In order to select the dimension p_0 of the true model, we need to look at the likelihood ratio test statistic (LRTS). The LRTS is defined as:

$$2\lambda_n = 2\left(\sup_{g \in \mathcal{G}} l_n(g) - l_n(g^0)\right) \tag{1}$$

For regular statistical models, the LRTS converges to a χ^2 distribution. This is no longer the case with this model. Let us first recall a result which gives an approximation of the LRTS.

2.1 Approximation of the LRTS

In the classical statistical theory, the approximation of the log-likelihood and the LRTS use a Taylor expansion in a neighborhood of the true parameter by computing the first derivatives and second derivatives of the log-likelihood, for example if the distribution g depends on a parameter θ, and the true parameter of g^0 is θ^0:

$$l_n(g) \simeq l_n(g^0) + (\theta - \theta^0)\frac{\partial l_n(g)}{\partial \theta}(\theta^0) + \frac{1}{2}(\theta - \theta^0)^2 \frac{\partial^2 l_n(g)}{\partial \theta^2}(\theta^0)$$

The first derivative $\frac{\partial l_n(g)}{\partial \theta}(\theta^0)$ is called the score function. However, our model is no more identifiable (the true parameter is not unique), so we can not use this expansion. Instead we define a new "extended set of score-functions" which allows us to give an approximation of the LRTS in a neighborhood of

the set of true parameters (the parameters giving the true log-likelihood function).

In order to get such approximation, we have to control the size of the term λ_n, which can be done thanks the "empirical processes theory" (see van der Vaart [15]). This theory deals with "law of large number" and "asymptotic normality" for set of functions. To get the law of large number, the considered set of function has to be not too big, we say "Glivenko-Cantelly" (see van der Vaart [15], page 269). This means that the set of function can be covered by a finite set of balls, i.e. the covering number (see below) of the set of function is finite. The assumptions for the "asymptotic normality" is more restrictive, now the covering number, which depends of the diameter ε of the balls, has to be of order $e^{\frac{1}{\varepsilon^2}}$ when ε goes to 0. We call such set a Donsker set (see van der Vaart [15], page 269).

Now, we have to introduce some definitions and properties.

- We recall that μ is the probability distribution of the vector (X_k, Y_k). For a function f such that $f(X_k, Y_k)$ is square integrable, let $\|f(X_k, Y_k)\|_{L^2(\mu)}$ be

$$\sqrt{\int f^2(X_k, Y_k) d\mu(X_k, Y_k)}$$

- For $\eta > 0$, let us denote $\mathcal{G}_\eta := \left\{ g \in \mathcal{G}, \| \frac{g}{g^0} - 1 \|_{L^2(\mu)} \leq \eta \right\}$, $\mathcal{G}_\eta$ is then the set of conditional density functions in a neighborhood of the true conditional density function g^0. In this neighborhood let us define the extended set of score-functions $\mathcal{S}_\eta$ as:

$$\mathcal{S}_\eta = \left\{ s_g = \frac{\frac{g}{g^0} - 1}{\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)}}, g \in \mathcal{G}_\eta \right\}.$$

- Consider the extended set of score-functions $\mathcal{S}_\eta$ endowed with the norm $\|\cdot\|_{L^2(\mu)}$. For every $\varepsilon > 0$, we define an ε-bracket by $[l, u] = \{ f \in \mathcal{S}_\eta, \, l \leq f \leq u \}$ such that $\|u - l\|_{L^2(\mu)} < \varepsilon$. The ε-bracketing entropy is

$$\mathcal{H}_{[\cdot]} \left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{L^2(\mu)} \right) = \ln \left(\mathcal{N}_{[\cdot]} \left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{L^2(\mu)} \right) \right),$$

where $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{L^2(\mu)}\right)$ is the minimum number of ε-brackets necessary to cover $\mathcal{S}_\eta$. $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{L^2(\mu)}\right)$ is also called "covering number".

- With the previous notations, we introduce the following assumption **(B)**: Assume that $\mathcal{G}$ is Glivenko-Cantelli and that there exists $\eta > 0$ such that
$$\int_0^1 \sqrt{\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{L^2(\mu)}\right)} d\varepsilon < \infty.$$
Then the set $\mathcal{S}_\eta$ is Donsker under **(B)**.

- Let us also define the limit-set of scores $\mathcal{D}$
$$\left\{ d \in \mathbb{L}^2(\mu) \mid \exists (g_n) \in \mathcal{G}, \|\tfrac{g_n - g^0}{g^0}\|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0, \|d - s_{g_n}\|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0 \right\}.$$
By putting $g_t = g_n$ for $t \in [0,1]$ and $n \leq \frac{1}{t} < n+1$, we obtain that, for all $d \in \mathcal{D}$, there exists a parametric path $(g_t)_{0\leq t\leq 1}$ such that $\forall t \in [0,1]$, $g_t \in \mathcal{G}$, $t \to \|\frac{g_t - g^0}{g^0}\|_{\mathbb{L}^2(\mu)}$ is continuous in 0, $\|\frac{g_t - g^0}{g^0}\|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$ and $\|d - s_{g_t}\|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$.

The following theorem can be stated (Gassiat, 2002):

Theorem 1: *Under the assumption* **(B)**,
$$2\lambda_n = \sup_{d\in\mathcal{D}} \left(\max\left\{ \tfrac{1}{\sqrt{n}} \sum_{i=2}^n d(Y_i, X_i); 0 \right\} \right)^2 + o_P(1)$$

Using this result, we may study a simple example of such model.

2.2 Simple mixture of experts

In this section, we shall test one against two components in the mixture of experts.
Let $\mathcal{G}$ be the set of possible conditional densities:

$$\mathcal{G} = \left\{ g(y \mid x) = \pi_\nu(x) g_\theta(y \mid x) + (1 - \pi_\nu(x)) g^0(y \mid x), \pi_\nu(x) \in [0;1], g_\theta \in \mathcal{P} \right\}$$

with $\mathcal{P} = \left\{ g_\theta(y|x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta x)^2}, \theta \in \Theta \subset \mathbb{R} \right\}$ the set of conditional densities and $g^0\left(y \mid x\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$. This model is clearly a particular case of the general mixture of expert model and is a simple example of mixture of regressions with Gaussian noise. Note that when it comes to the decision about the number of components, if the true conditional density function is in set set of possible parameter, say $\mathcal{G}^*$, then $\mathcal{G} \subset \mathcal{G}^*$ and

$$2\left(\sup_{g\in\mathcal{G}^*} l_n(g) - l_n(g^0)\right) \geq 2\left(\sup_{g\in\mathcal{G}} l_n(g) - l_n(g^0)\right)$$

So, if the behavior of λ_n is bad for $g \in \mathcal{G}$, it is worst for a more general case.

Within this frame, the test may be rewritten as $\theta \neq 0$ and $\pi_\nu(x) \neq 0$ (two components) against $\theta = 0$ and $\pi_\nu(x) = 0\ ,\ \forall x$ (one component).
The LRTS is defined as:

$$2\lambda_n = 2\left(\sup_{g\in\mathcal{G}} \ln(g) - \ln(g^0)\right) = 2\sup_{g\in\mathcal{G}} \sum_{k=1}^{n} \ln \frac{\pi g_\theta\left(Y_k \mid X_k\right) + (1-\pi) g^0\left(Y_k \mid X_k\right)}{g^0\left(Y_k \mid X_k\right)} \tag{2}$$

In order to derive the behavior of the LRTS, two cases have to be analyzed. The first one is if there exists a sequence of parameters $\nu_1, \cdots, \nu_k$ such that $\lim_{k\to\infty} E\left[\pi_{\nu_k}(X)\right] = 0$. The second one is when $\exists \delta > 0$ such that $\forall \nu, E\left[\pi_\nu(X)\right] \geq \delta$.

2.2.1 Divergence of LRTS

The LRTS can be divergent if there exists a sequence of parameters $\nu_1, \cdots, \nu_k, \cdots$ such that $\lim_{k\to\infty} E\left[\pi_{\nu_k}(X)\right] = 0$. Indeed, for such sequence we can have $\| \ln(g) - \ln(g^0) \| \to 0$ with $\theta \neq 0$.

For sake of simplicity, assume that the probability function $\pi_\nu(X)$ is constant. Then, the corresponding score functions are :

$$\frac{\frac{g}{g^0} - 1}{\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)}} = \frac{\exp\left(-\frac{\theta^2}{2}X^2 + \theta Y X\right) - 1}{\left\| \exp\left(-\frac{\theta^2}{2}X^2 + \theta Y X\right) - 1 \right\|_{L^2(\mu)}} \tag{3}$$

and if the quantity
$\left\| \exp\left(-\frac{\theta^2}{2}X^2 + \theta Y X\right) - 1 \right\|_{L^2(\mu)}$ is finite, the score functions are well defined.

Let us study

$$\left\|\exp\left(-\frac{\theta^2}{2}X^2+\theta YX\right)-1\right\|^2_{L^2(\mu)}=$$
$$\frac{1}{2\pi}\int\int\left(\exp\left(-\frac{\theta^2}{2}x^2+\theta yx\right)-1\right)^2\exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}y^2\right)dxdy=$$
$$\frac{1}{2\pi}\int\int\left(\exp\left(-\theta^2x^2+2\theta yx\right)-2\exp\left(-\frac{\theta^2}{2}x^2+\theta yx\right)+1\right)$$
$$\exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}y^2\right)dxdy$$

The integral of the dominant term (the first) is:

$$\begin{aligned}I(\theta) &= \frac{1}{2\pi}\int\int\exp\left(-\theta^2x^2+2\theta yx\right)\exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}y^2\right)dxdy\\ &= \frac{1}{2\pi}\int\int\exp\left(-\left(\theta^2+\frac{1}{2}\right)x^2+2\theta xy-\frac{1}{2}y^2\right)dxdy\\ &= \int\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\left(1-2\theta^2\right)\right)\left(\int\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(y-\theta x\right)^2\right)dy\right)dx\\ &= \frac{1}{\sqrt{1-2\theta^2}}\end{aligned}$$

Thus, for $-\frac{1}{\sqrt{2}}<\theta<\frac{1}{\sqrt{2}}$, $\left\|\exp\left(-\frac{\theta^2}{2}X^2+\theta YX\right)-1\right\|_{L^2(\mu)}<+\infty$ and the score function is well defined. The set of limit score functions contains the score functions:

$$\left\{s_\theta(X,Y)=\frac{\frac{g_\theta(Y|X)}{g^0(Y|X)}}{\|\frac{g_\theta(Y|X)}{g^0(Y|X)}\|_{L^2(\mu)}},\theta\in\left]-\frac{1}{\sqrt{2}};\frac{1}{\sqrt{2}}\right[\right\}$$

Suppose that an arbitrary number of "almost" uncorrelated random variables in C can be found, then λ_n can take an arbitrarily large value since the maximum of m independent samples from standard normal distribution is approximately $\sqrt{2\log m}$. According to Fukumizu (2003), if a sequence $\theta_1,\cdots,\theta_m,\cdots$ exists so that

$$\lim_{m\to\infty}s_{\theta_m}(X,Y)\xrightarrow{P}0$$

then the likelihood ratio diverges to infinity. Here, we get

$$\lim_{\theta\to\frac{1}{\sqrt{2}},\theta<\frac{1}{\sqrt{2}}}\left\|\exp\left(-\frac{\theta^2}{2}X^2+\theta YX\right)-1\right\|_{L^2(\mu)}=+\infty$$

So, for each sphere B of $\mathbb{R}^2$, centered on the origin, if $(x,y)\in B$:

$$\lim_{\theta\to\frac{1}{\sqrt{2}},\theta<\frac{1}{\sqrt{2}}}\frac{\exp\left(-\frac{\theta^2}{2}x^2+\theta yx\right)-1}{\left\|\exp\left(-\frac{\theta^2}{2}X^2+\theta YX\right)-1\right\|_{L^2(\mu)}}=0$$

and $\frac{\exp\left(-\frac{\theta^2}{2}X^2+\theta YX\right)-1}{\left\|\exp\left(-\frac{\theta^2}{2}X^2+\theta YX\right)-1\right\|_{L^2(\mu)}}$ converges to 0 in probability for $\theta\to\frac{1}{\sqrt{2}},\theta<\frac{1}{\sqrt{2}}$. With the choice $\theta_m=\frac{1}{\sqrt{2}}-\frac{1}{m}$, we get $\lim_{m\to\infty}s_{\theta_m}(x,y)\xrightarrow{P}0$ and the LRTS is divergent.

2.2.2 Convergence of LRTS

In this section we suppose that $(\exists)\delta > 0$ such that $(\forall)\nu$, $\mathbb{E}(\pi_\nu(X)) \geq \delta > 0$. Since $\pi_\nu(X) \geq 0$, then $(\exists)A \subseteq \mathbb{R}$ such that $\lambda(A) = \eta > 0$ and $\pi_\nu(x) > \delta$ for any $x \in A$. In such case, the maximum likelihood estimator $\hat{\theta}$ converges to $\theta_0 = 0$, otherwise $\lim_{n\to\infty} \lambda_n = \sup_{g\in\mathcal{G}} E_\mu(\ln(g) - \ln(g^0))$ can not be close to 0.

Here, the generalized score-function can be rewritten using the following:

$$\begin{aligned} \frac{g}{g^0} - 1 &= \frac{\pi_\nu(X)g_\theta(Y|X)+(1-\pi_\nu(X))g^0(Y|X)}{g^0(Y|X)} - 1 \\ &= \pi_\nu(X)\left(\frac{g_\theta(Y|X)}{g^0(Y|X)} - 1\right) \\ s_g = \frac{\frac{g}{g^0}-1}{\|\frac{g}{g^0}-1\|_{L^2(\mu)}} &= \frac{\pi_\nu(X)\left(\frac{g_\theta(Y|X)}{g^0(Y|X)}-1\right)}{\|\pi_\nu(X)\left(\frac{g_\theta(Y|X)}{g^0(Y|X)}-1\right)\|_{L^2(\mu)}} \end{aligned}$$

The model is parameterized by $\phi = (\theta, \nu) \in \Theta \times V \subseteq \mathbb{R}^2$ compact set and θ_0 belongs to the interior of Θ. Since $\mathbb{E}(\pi_\nu(X)) \geq \delta > 0$, we have that $g = g_0 \Leftrightarrow \theta = \theta_0$. Thus, the model is identifiable in θ and unidentifiable in ν. For any fixed $\nu \in V$, we have the following Taylor expansion around θ_0:

$$l_{(\theta,\nu)} - 1 = (\theta - \theta_0)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)} + o(|\theta - \theta_0|)$$

where $l_{(\theta,\nu)} = \frac{g_\theta}{g^0}$. Hence,

$$\begin{aligned} s_g = s_{\phi=(\theta,\nu)} &= \frac{\frac{g}{g^0}-1}{\|\frac{g}{g^0}-1\|_{L^2(\mu)}} \\ &= \frac{\pi_\nu(X)\left[(\theta-\theta_0)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(|\theta-\theta_0|)\right]}{\|\pi_\nu(X)\left[(\theta-\theta_0)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(|\theta-\theta_0|)\right]\|_{L^2(\mu)}} \\ &= \beta\frac{\pi_\nu(X)\left[\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(1)\right]}{\|\pi_\nu(X)\left[\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(1)\right]\|_{L^2(\mu)}} \end{aligned}$$

where $|\beta| = 1$.

In the Gaussian case, $g_\theta(y|x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\theta x)^2\right)$, the first derivative of $l_{(\theta,\nu)}$ is $\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}(x,y) = xy$, hence the directional score functions are not linearly correlated ($\beta\pi_\nu(x)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}(x,y) \neq 0$ for any β and $\nu \in V$) and we may apply Lemma 4.1 in Liu and Shao [13]. According to this lemma, when $|\theta - \theta_0| \to 0$, the set of limit score functions is $\mathcal{F}$, where

$$\mathcal{F} = \left\{\Omega\left(\beta\pi_\nu(X)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}(X,Y)\right), |\beta| = 1, \nu \in V\right\}$$

and $\Omega(f) = \frac{f}{\|f\|_{L^2(\mu)}}$. According to Theorem 3.1 in Liu and Shao [13], the LRTS satisfies:

$$\lim 2\lambda_n = \sup_{s_g \in \mathcal{F}} \left(W_{s_g} \vee 0\right)^2$$

where $\left\{W_{s_g}, s_g \in \mathcal{F}\right\}$ is a centered Gaussian process with continuous sample paths and covariance kernel $\mathbb{E}\left(W_{s_1} W_{s_2}\right) = \mathbb{E}\left(s_1 s_2\right)$. In our case,

$$\mathcal{F} = \left\{\Omega\left(\beta \pi_\nu(X) XY\right), |\beta| = 1, \nu \in V\right\}$$

and

$$\mathbb{E}\left(s_1 s_2\right) = \frac{\mathbb{E}\left(X^2 Y^2 \pi_{\nu_1}(X) \pi_{\nu_2}(X)\right)}{\|\pi_{\nu_1}(X) XY\|_{L^2(\mu)} \|\pi_{\nu_2}(X) XY\|_{L^2(\mu)}}$$

3 General case

This section describes the asymptotics of the LRTS in the general frame introduced in the beginning of the previous section. We shall prove that assumption **(B)** holds for mixtures of experts (ME) under some general hypothesis. Furthermore, we shall prove that the limit set of scores $\mathcal{D}$ is complete and has continuous parametric paths. Hence, the asymptotic behavior of the LRTS may be completely explicated.

Assumptions for the tightness of LRTS

H-1 The set $\mathcal{G}$ is Glivenko-Cantelli and the set of possible parameters contains a neighborhood of the parameters defining the true conditional density g^0.

H-2 There exists $\eta > 0$ such that for all $g \in \mathcal{G}$ with $\|g - g^0\|_{L^2(\mu)} \leq \eta$, $\left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)} < \infty$

H-3 By denoting $l_{\theta_i} := \frac{g_{\theta_i}}{g^0}$ and, with a slight abuse of notation, $\frac{\partial^q}{\partial \theta_j^q}$ the derivative of order q with respect to all components of θ_j, we assume the existence of a square-integrable function h such that, for all $(\theta_1, \cdots, \theta_p)$,

$$\left|\frac{\partial l_{\theta_j}}{\partial \theta_j}(\theta_j)\right| \leq h, \ \left|\frac{\partial^2 l_{\theta_j}}{\partial \theta_j^2}(\theta_j)\right| \leq h \text{ and } \left|\frac{\partial^3 l_{\theta_j}}{\partial \theta_j^3}(\theta_j)\right| \leq h.$$

H-4 With the following notations:

$$l'_j := \frac{\partial l_{\theta_j}}{\partial \theta_j}\left(\theta_j^0\right),\ l'' := \frac{\partial^2 l_{\theta_j}}{\partial \theta_j^2}\left(\theta_j^0\right)$$

we assume that for distinct $(\theta_i)_{1\leq i\leq p}$

$$\left\{(l_{\theta_i})_{1\leq i\leq p}, (l'_i)_{1\leq i\leq p^0}, (l''_i)_{1\leq i\leq p^0}\right\}$$

are linearly independent in the Hilbert space $L^2(\mu)$.

The key assumption **(H-2)** guarantees that the score functions are well defined. This assumption failed in the example of divergence of the LRTS. In the general case, one way to get this assumption to hold is to have the expected values of mixture weights $E\|\pi_\nu(X)\|_{L^2(\mu)}$ greater than some positive constant (see the previous example).

Let us define $\Omega : L^2(P) \to L^2(\mu)$ by $\Omega(g) = \frac{g}{\|g\|_2}$, for $g \neq 0$.

We can state the following theorem, which generalizes theorem 4.1 of Liu and Shao [13] :

Theorem 2:

*Let D be the parametric dimension of the regression functions. Under the assumptions **(H-1)**, **(H-2)**, **(H-3)** and **(H-4)**, there exists a centered Gaussian process $\{W_S, S \in \mathbb{F}\}$ with continuous sample path and covariance kernel $P(W_{S_1}W_{S_2}) = P(S_1S_2)$ such that*

$$\lim_{n\to\infty} 2\lambda_n = \sup_{S\in\mathbb{F}} \left(\max(W_S, 0)\right)^2.$$

The index set $\mathbb{F}$ is defined as $\mathbb{F} = \cup_t \mathbb{F}_t$, with the union running over $t = (t_0, \cdots, t_{p_0}) \in \mathbb{N}^{p_0+1}$ with $0 = t_0 < t_1 < \cdots < t_{p_0} \leq p$ and

$$\begin{array}{l}\mathbb{F}_t = \Big\{\Omega\Big(\sum_{i=1}^{p_0}\zeta_i(X)l_{\theta_i^0} + \sum_{i=p_0+1}^{p}\zeta_i(X)l_{\theta_i} + \\ \sum_{i=1}^{p_0}\lambda_i^T(X)l'_i + \delta\sum_{i=1}^{p_0}\sum_{j=t_{i-1}+1}^{t_i}\gamma_j^T(X)l''_i\gamma_j(X)\Big), \\ \text{for all } x,\ \lambda_1(x), \cdots, \lambda_{p_0}(x), \gamma_1(x), \cdots, \gamma_{t_{p_0}}(x) \in \mathbb{R}^d\ ; \zeta_1(x), \cdots, \zeta_p(x) \in \mathbb{R}, \\ \theta_{t_{p_0}+1}, \cdots, \theta_p \in \Theta - \left\{\theta_1^0, \cdots, \theta_{p_0}^0\right\}\Big\}\end{array}$$

where $\delta = 1$ if there exists a vector $\mathbf{q}$ such that:
$q_j(x) \geq 0$, $\sum_{j=t_{i-1}+1}^{t_i} q_j(x) = 1$, $\sum_{j=t_{i-1}+1}^{t_i}\sqrt{q_j(x)}\gamma_j^t(x) = 0$ for $i = 1, \cdots, p_0$; and $\delta = 0$ otherwise.

Note that the asymptotic law of the LRTS depends on the true parameters of the model. The proof of this theorem is postponed in the appendix.

4 Conclusion

Summary of the findings Mixtures of Experts and Hierarchical Mixtures of Experts are powerful tools to deal with regression models, maybe too powerful, indeed, for models like :

$$g\left(y|x,\phi\right)=\sum_{i=1}^{p}\pi_{\nu_i}(x)g_{\theta_i}(y|x),$$

The divergence of the LRTS shows that strong overfitting problem plagues theses models. But if we consider the restricted models :

$$g\left(y|x,\phi\right)=\sum_{i=1}^{p}\pi_{\nu_i}(x)g_{\theta_i}(y|x),\ \text{with}\ \pi_{\nu_i}(x)\geq\eta$$

for a small constant $\eta > 0$ the theorem 2 shows that the overfitting will be moderated. This conditions means that the user can introduce new regression in the mixture density if the probability of the new regression is not too small. Practically, $\eta = 0.01$ or $\eta = 0.001$ seems to be a reasonable choice.

As a conclusion The example of this paper illustrates the two main behaviors that one can expect: moderate overfitting if the mixing probabilities are bounded from below and strong overfitting if the mixing probabilities can be as small as possible. Moreover, the main interest of the general theorem 2 is to show that the LRTS is tight under some general assumptions, such that the true number of components of the mixture may be selected thanks to classical penalized log-likelihood criteria like BIC. So, if the user seeks to minimize

$$l_n\left(g\right)+D\times\log(n)$$

where D is the number of parameter of the models, then it will automatically select the true number of components of the mixture of expert if n, the number of observations, is large enough.

Appendix : Proof of Theorem 2

Let $\eta > 0$ be a real number. Consider $\hat{\mathcal{G}}_n \neq \emptyset$ the set of functions which maximize the log-likelihood. Since, under **(H-1)**, $\mathcal{G}$ is Glivenko-Cantelli, for

n large enough, $\|g - g^0\|_{L^2(\mu)} < \eta$ for $g \in \hat{\mathcal{G}}_n$ so $\hat{\mathcal{G}}_n \subset \mathcal{G}_\eta$. Let us remark that, under assumption **(H-2)**, the score function $s_g \in \mathcal{S}_\eta$ is well defined in a compact neighborhood of the true density function g^0.

Proving that for an $\eta > 0$, a parametric family like $\mathcal{S}_\eta$ is Donsker is not straightforward. The problems arise when $g \to g^0$ and the limits of s_g in $L^2(\mu)$ have to be computed. To achieve our proof, let us split $\mathcal{S}$ into two classes of functions.

For a sufficiently small $\varepsilon > 0$, we consider $\mathcal{F}_0 \subset \mathcal{G}_\eta$, a neighborhood of g^0, $\mathcal{F}_0 = \left\{ g \in G,\ \left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)} \leq \varepsilon,\ g \neq g^0 \right\}$. $\mathcal{S}_\eta$ is split into $\mathcal{S}_0 = \{ s_g,\ g \in \mathcal{F}_0 \}$ and $\mathcal{S}_\eta \setminus \mathcal{S}_0$.

On $\mathcal{S}_\eta \setminus \mathcal{S}_0$, it can be easily seen that

$$\left\| \frac{\frac{g_1}{g^0} - 1}{\left\| \frac{g_1}{g^0} - 1 \right\|_{L^2(\mu)}} - \frac{\frac{g_2}{g^0} - 1}{\left\| \frac{g_2}{g^0} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} \leq 2 \frac{\left\| \frac{g_1}{g^0} - \frac{g_2}{g^0} \right\|_{L^2(\mu)}}{\left\| \frac{g_1}{g^0} - 1 \right\|_{L^2(\mu)}}$$

for every $g_1, g_2 \in \mathcal{G}_\eta \setminus \mathcal{F}_0$ and, moreover, by the definition of $\mathcal{S}_0$,

$$\left\| \frac{\frac{g_1}{g^0} - 1}{\left\| \frac{g_1}{g^0} - 1 \right\|_{L^2(\mu)}} - \frac{\frac{g_2}{g^0} - 1}{\left\| \frac{g_2}{g^0} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} \leq \frac{2}{\varepsilon} \left\| \frac{g_1}{g^0} - \frac{g_2}{g^0} \right\|_{L^2(\mu)}$$

On the other hand, by the assumption **(H-3)**, $\frac{g}{g^0}$ has square-integrable partial-derivatives of order one and, using the result 19.7 on parametric classes of functions in [15], we get:

$$\mathcal{N}_{[\cdot]} \left(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2 \right) = \mathcal{O} \left(\frac{1}{\varepsilon^2} \right)^D,$$

where D is the number of parameters in the model.

It remains to prove that the bracketing number is a polynomial of $(\frac{1}{\varepsilon})$ for $\mathcal{S}_0$. The idea is to reparameterize the model in a convenient manner which will allow a Taylor expansion around the identifiable part of the true value of the parameters.

Let us recall that it is assumed that $p_0 < p$.

When $\frac{g}{g^0} - 1 = 0$, by the linear independence of the functions g_{θ_j}, a vector of positive integers $t = (t_i)_{0 \leq i \leq p_0}$, $t_0 = 0$ exists so that:

$$\theta_{t_{i-1}+1} = ... = \theta_{t_i} = \theta_i^0, \ \sum_{j=t_{i-1}+1}^{t_i} \pi_{\nu_j}(x) = \pi_{\nu_i}^0(x), \ i \in \{1, ..., p_0\}$$

With this remark, one can define in the general case $s(x) = (s_i(x))_{1 \leq i \leq p_0}$ and $q(x) = (q_j(x))_{1 \leq j \leq p}$ so that, for every $i \in \{1, ..., p_0\}$, $j \in \{t_{i-1} + 1, ..., t_i\}$,

$$s_i(x) = \sum_{j=t_{i-1}+1}^{t_i} \pi_{\nu_j}(x) - \pi_{\nu_i}^0(x), \ q_j(x) = \frac{\pi_{\nu_j}(x)}{\sum_{l=t_{i-1}+1}^{t_i} \pi_{\nu_l}(x)}$$

and a new parameterization will be

$$\Theta_t = (\phi_t, \psi_t) \, , \ \phi_t = \left((\theta_j)_{1 \leq j \leq t_{p_0}} \, , (s_i(x))_{1 \leq i \leq p_0 - 1} \, , \left(\pi_{\nu_j}(x)\right)_{j=t_{p_0}+1}^{p} \right) ,$$
$$\psi_t = \left((q_j(x))_{1 \leq j \leq p} \, , (\theta_j)_{j=t_{p_0}+1}^{p} \right)$$

with ϕ_t containing all the identifiable parameters of the model and ψ_t the non-identifiable ones. Then, for $g = g^0$, we will have:

$$\phi_t^0 = \ (\underbrace{\theta_1^0, ..., \theta_1^0}_{t_1} \ \ , ..., \ \underbrace{\theta_{p_0}^0, ..., \theta_{p_0}^0}_{t_{p_0} - t_{p_0 - 1}}, \ \underbrace{0, ..., 0}_{p_0 - 1} \ \underbrace{0, ..., 0}_{p - t_{p_0}} \ )^T$$

This reparameterization allows to write a second-order Taylor expansion of $\frac{g}{g^0} - 1$ at ϕ_t^0.

With the notations introduced in assumptions **(H1)-(H4)**, the density ratio becomes:

$$\frac{g}{g^0} - 1 = \sum_{i=1}^{p_0} \left(s_i(x) + \pi_{\nu_i}^0(x) \right) \sum_{j=t_{i-1}+1}^{t_i} q_j(x) l_{\theta_j} + \sum_{j=t_{p_0}+1}^{p} \pi_{\nu_j}(x) l_{\theta_j} - 1$$

and since $s_{p_0}(x) = - \sum_{i=1}^{p_0 - 1} s_i(x)$,

$$\begin{aligned} \frac{g}{g^0} - 1 = \ & \textstyle\sum_{i=1}^{p_0-1} \left(s_i(x) + \pi_{\nu_i}^0(x) \right) \sum_{j=t_{i-1}+1}^{t_i} q_j(x) l_{\theta_j} + \\ & \left(\pi_{\nu_{p_0}}^0 - \textstyle\sum_{i=1}^{p_0-1} s_i(x) \right) \sum_{j=t_{p_0-1}+1}^{t_{p_0}} q_j(x) l_{\theta_j} \\ & + \textstyle\sum_{j=t_{p_0}+1}^{p} \pi_{\nu_j}(x) l_{\theta_j} - 1 \end{aligned}$$

By remarking that when $\phi_t = \phi_t^0$, $\frac{g}{g^0}$ does not vary with ψ_t, we will study the variation of this ratio in a neighborhood of ϕ_t^0 and for fixed ψ_t.

We can state the following result:. The proof is an easy application of Taylor's formula and will be omitted.

Proposition 3

Let us denote $D(\phi_t, \psi_t) = \left\| \frac{g_{(\phi_t,\psi_t)}}{g^0} - 1 \right\|_{L^2(\mu)}$. *With the notations of assumptions* ***(H-3)*** *and* ***(H-4)****, for any fixed* ψ_t*, the second-order Taylor expansion at* ϕ_t^0 *exists such as*

$$\frac{g}{g^0} - 1 = \left(\phi_t - \phi_t^0\right)^T l'_{\left(\phi_t^0,\psi_t\right)} + \frac{1}{2}\left(\phi_t - \phi_t^0\right)^T l''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t - \phi_t^0\right) - 1 + o\left(D\left(\phi_t, \psi_t\right)\right)$$

with

$$\left(\phi_t - \phi_t^0\right)^T l'_{\left(\phi_t^0,\psi_t\right)} = \sum_{i=1}^{p_0} \pi_{\nu_i}^0(x) \left(\sum_{j=t_{i-1}+1}^{t_i} q_j(x)\theta_j - \theta_i^0 \right)^T l'_i + \sum_{i=1}^{p_0} s_i(x) l_{\theta_i^0} + \sum_{j=t_{p_0}+1}^{p} \pi_{\nu_j}(x) l_{\theta_j}$$

and

$$\left(\phi_t - \phi_t^0\right)^T l''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t - \phi_t^0\right) = \sum_{i=1}^{p_0} \left[2s_i(x) \left(\sum_{j=t_{i-1}+1}^{t_i} q_j(x)\theta_j - \theta_i^0 \right)^T l'_i + \pi_{\nu_i}^0(x) \sum_{j=t_{i-1}+1}^{t_i} q_j(x) \left(\theta_j - \theta_i^0\right)^T l''_i \left(\theta_j - \theta_i^0\right) \right]$$

Moreover,

$$\left(\phi_t - \phi_t^0\right)^T l'_{\left(\phi_t^0,\psi_t\right)} + \frac{1}{2}\left(\phi_t - \phi_t^0\right)^T l''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t - \phi_t^0\right) = 0 \Leftrightarrow \phi_t = \phi_t^0$$

Using the Taylor expansion above, we can now show that $\mathcal{S}_0 \setminus \{g^0\}$ is a Donsker class, using the next result:

Proposition 4

Let d be the dimension of the parameter indexing the functions g_θ. The number of ε-brackets $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_0, \|\cdot\|_2\right)$ covering $\mathcal{S}_0$ is $\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{p_0\times(2d)+p}$.

Proof of Proposition 4

The idea of this proof is to bound $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_0, \|\cdot\|_2\right)$ by the number of ε-brackets covering a wider class of functions. For every $g \in \mathcal{F}_0$, we will consider the reparameterization $\Phi = (\phi_t, \psi_t)$ which allows to write a second-order development of the density ratio:

$$\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1 = \left(\phi_t - \phi_t^0\right)^T l'_{\left(\phi_t^0,\psi_t\right)} + \frac{1}{2}\left(\phi_t - \phi_t^0\right)^T l''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t - \phi_t^0\right) + o\left(D\left(\phi_t, \psi_t\right)\right)$$

Then, by remarking that the first two terms in the Taylor expansion are linear combinations of $l_{\theta_i^0}$, l'_i, l''_i, $i = 1, ..., p_0$ and $l_{\theta_j}, j = t_{p_0} + 1, \cdots, p$, the density ratio can be written also as:

$$\begin{array}{l}\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1 = \\ \sum_{i=1}^{p_0} \alpha_i(X) l_{\theta_i^0} + \sum_{j=t_{p_0}+1}^{p} \alpha_j(X) l_{\theta_j} + \sum_{i=1}^{p_0} \beta_i^T(X) l'_i \\ + \sum_{i=1}^{p_0} \gamma_i^T(X) l''_i \gamma_i(X) + o\left(D\left(\phi_t, \psi_t\right)\right)\end{array}$$

where, for all x, $(\alpha_i(x))_{1\le i\le p} \in \mathbb{R}$, $(\beta_i(x))_{1\le i\le p_0}$ and $(\gamma_i(x))_{1\le i\le p_0} \in \mathbb{R}^d$.

Now, using the linear independence, $\exists m > 0$,so that, for every

$$\left(\alpha_j(x), j = 1, \cdots, p, \beta_i(x), \gamma_i(x)\gamma_i^T(x), i = 1, \cdots, p_0\right)$$

of norm 1,

$$\left\| \sum_{i=1}^{p_0} \alpha_i(X) l_{\theta_i^0} + \sum_{j=t_{p_0}+1}^{p} \alpha_j(X) l_{\theta_j} + \sum_{i=1}^{p_0} \beta_i^T(X) l'_i + \sum_{i=1}^{p_0} \gamma_i^T(X) l''_i \gamma_i(X) \right\|_{L^2(\mu)} \ge m.$$

At the same time, since

$$\left\| \frac{\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1}{\left\| \frac{g_{(\phi_t,\psi_t)}}{g^0} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} = 1$$

we will obtain that the Euclidean norm of the coefficients in the second-order development of $\frac{\frac{g_{(\phi_t,\psi_t)}}{g^0}-1}{\left\|\frac{g_{(\phi_t,\psi_t)}}{g^0}-1\right\|_{L^2(\mu)}}$ is upper bounded by $\frac{1}{m}$. This fact implies that $\mathcal{S}_0$ can be included in

$$\mathcal{H} = \begin{array}{l} \left\{ \sum_{i=1}^{p_0} \left(\alpha_i(x) l_{\theta_i^0} + \beta_i^T(x) l_i' + \gamma_i^T(x) l_i'' \gamma_i(x) \right) + \sum_{j=t_{p_0}+1}^{p} \alpha_j(x) l_{\theta_j} + o(1), \right. \\ \left. \left\| \left(\alpha_j(x), j = t_{p_0}+1, \cdots, p, \beta_i(x), \gamma_i(x)\gamma_i^T(x), i = 1, \cdots, p_0 \right) \right\| \leq \frac{1}{m} \right\} \end{array}$$

and then obviously $\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{H}, \|\cdot\|_2\right) = \mathcal{O}\left(\frac{1}{\varepsilon}\right)^{p_0 \times 2d + p + 1}$. ■

Since the set $\mathcal{S}_\eta$ was proven to be Donsker, it remains to identify the asymptotic index set of score functions.

Asymptotic index set. The set of limit score functions $\mathbb{F}$ is defined as the set of functions d so that one can find a sequence g_n satisfying $\|\frac{g_n - f}{f}\|_2 \to 0$ and $\|d - s_{g_n}\|_2 \to 0$.

Let us define the two principal behaviors for the sequences g_n which influence the form of functions d :

- If the second order term is negligible with respect to the first one :

$$\frac{g_n}{g^0} - 1 = (\Phi_n - \Phi^0)^T l'_{(\Phi_t^0, \psi_n)} + o(D(\Phi_n, \psi_n)).$$

- If the second order term is not negligible with respect to the first one :

$$\begin{array}{l} \frac{g_n}{g^0} - 1 = (\Phi_n - \Phi^0)^T l'_{(\Phi_t^0, \psi_n)} + \\ 0.5(\Phi_n - \Phi^0)^T l''_{(\Phi^0, \psi_n)} (\Phi_n - \Phi^0) + o(D(\Phi_n, \psi_n)). \end{array}$$

In the first case, a set $t = (t_0, \cdots, t_{p_0})$ exists so that the limit function of s_{g_n} will be in the set:

$$\begin{array}{ll} \mathbb{D}_1^t = & \left\{ \Omega \left(\sum_{i=1}^{p_0} \zeta_i^T(X) l_{\theta_i^0} + \sum_{i=p_0+1}^{p} \zeta_i^T(X) l_{\theta_i} + \sum_{i=1}^{p_0} \lambda_i^T(X) l_i' \right) \right. \\ \text{for all } x & \lambda_1(x), \cdots, \lambda_{p_0}(x) \in \mathbb{R}^d \, ; \zeta_1(x), \cdots, \zeta_p(x) \in \mathbb{R} \\ & \left. \theta_{t_{p_0}+1}, \cdots, \theta_p \in \Theta - \left\{ \theta_1^0, \cdots, \theta_{p_0}^0 \right\} \right\} \end{array}$$

In the second case, an index i exists so that :

$$\sum_{j=t_{i-1}+1}^{t_i} q_j(x)(\theta_j - \theta_i^0) = 0,$$

Otherwise, the second order term will be negligible compared to the first one, so

$$\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j(x)} \times \sqrt{q_j(x)}(\theta_j - \theta_i^0) = 0.$$

Hence, a set a set $t = (t_0, \cdots, t_{p_0})$ exists so that the set of functions d will be:

$\Big\{\Omega \Big(\sum_{i=1}^{p_0} \zeta_i(X) l_{\theta_i^0} + \sum_{i=p_0+1}^{p} \zeta_i(X) l_{\theta_i} + \sum_{i=1}^{p_0} \lambda_i^T(X) l_i'$
$+\delta \sum_{i=1}^{p_0} \sum_{j=t_{i-1}+1}^{t_i} \gamma_j^T(X) l_i'' \gamma_j(X)\Big)$
for all x, $\lambda_1(x), \cdots, \lambda_{p_0}(x), \gamma_1(x), \cdots, \gamma_{t_{p_0}}(x) \in \mathbb{R}^d$; $\zeta_1(x), \cdots, \zeta_p(x) \in \mathbb{R}$
$\theta_{t_{p_0}+1}, \cdots, \theta_p \in \Theta - \big\{\theta_1^0, \cdots, \theta_{p_0}^0\big\}\Big\}$

where $\delta = 1$ if there exists a vector $\mathbf{q}(x)$ exists so that:
$q_j(x) \geq 0$, $\sum_{j=t_{i-1}+1}^{t_i} q_j(x) = 1$, $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j(x)} \gamma_j^t(x) = 0$ for $i = 1, \cdots, p_0$; and $\delta = 0$ otherwise.

So, the limit functions will belong to $\mathbb{F}$. Conversely, let d be an element of $\mathbb{F}$, as functions d belong to the Hilbert sphere, one of their components is not equal to 0. Let us assume that this component is $\zeta_1(x)$, but the proof would be similar with any other component. The norm of d is 1, so any component of d is determined by the ratio: $\frac{\zeta_2(x)}{\zeta_1(x)}, \cdots, \frac{1}{\zeta_1(x)}\gamma_{p_0}(x)$.

Then, by assumption **(H-1)**, the set of possible parameters contains a neighborhood of the parameters realizing the true conditional density function g^0, we can chose the parameters of g_n so that:

$$\begin{aligned}
&\forall i \in \{2, \cdots, p_0\} \ : \ \frac{\sum_{j=t_{i-1}+1}^{t_i} \pi_{\nu_j}^n(x) - \pi_{\nu_i}^0(x)}{\sum_{j=1}^{t_1} \pi_{\nu_j}^n(x) - \pi_{\nu_1}^0(x)} \overset{n\to\infty}{\longrightarrow} \frac{\zeta_i(x)}{\zeta_1(x)},\\
&\forall i \in \{1, \cdots, p_0\} \ : \ \frac{\sum_{j=t_{i-1}+1}^{t_i} q_j^n(x)\big(\theta_j^n - \theta_i^0\big)}{\sum_{j=1}^{t_1} \pi_{\nu_j}^n(x) - \pi_{\nu_1}^0(x)} \overset{n\to\infty}{\longrightarrow} \frac{1}{\zeta_1(x)}\lambda_i(x),\\
&\forall j \in \{1, \cdots, t_{p_0}\} \ : \ \frac{\sqrt{q_j^n(x)}}{\sum_{j=1}^{t_1} \pi_{\nu_j}^n(x) - \pi_{\nu_1}^0(x)} \big(\theta_j^n - \theta_i^0\big) \overset{n\to\infty}{\longrightarrow} \frac{1}{\zeta_1(x)}\gamma_j(x),\\
&\forall i \in \{p_0 + 1, \cdots, p\} \ : \ \frac{\pi_{\nu_i}^n(x)}{\sum_{j=1}^{t_1} \pi_{\nu_j}^n(x) - \pi_{\nu_1}^0(x)} \overset{n\to\infty}{\longrightarrow} \frac{1}{\zeta_1(x)}\zeta_i(x).
\end{aligned}$$

■

References

[1] Dacunha-Castelle D., Gassiat E. (1997a) The estimation of the order of a mixture model, *Bernoulli*, **3**, 279-299

[2] Dacunha-Castelle D., Gassiat E. (1997b) Testing in locally conic models, *ESAIM Prob. and Stat.*, **1**, 285-317

[3] Dacunha-Castelle D., Gassiat E. (1999) Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes, *The Annals of Statistics*, **27(4)** , 1178-1209

[4] Fukumizu K. (2003) Likelihood ratio of unidentifiable models and multilayer neural networks, *Ann. Statist.* , **31**, 833-851

[5] Gassiat E., Keribin C. (2000) The likelihood ratio test for the number of components in a mixture with Markov regime, *ESAIM P&S*, **4**, 25-52

[6] Gassiat E. (2002) Likelihood ratio inequalities with applications to various mixtures, *Ann. Inst. Henri Poincare*, **38**, 897-906

[7] Henna J. (1985) On estimating the number of constituents of a finite mixture of continuous distributions, *Ann. Inst. Statist. Math.*, **37**, 235-240

[8] Izenman A.J., Sommer C. (1988) Philatelic mixtures and multivariate densities, *Journal of the American Stat. Assoc.*, **83**, 941-953

[9] Jacobs R.A., Jordan M.I., Nowlan S.J., Hinton G.E. (1991) Adaptive mixtures of local experts, *Neural Comp.*, **3**, 79-87

[10] Jiang W., Tanner M.A. (1999) On the asymptotic normality of Hierarchical Mixtures-of-Experts for Generalized Linear Models, *IEEE Trans. on Information Theory*, **46**, 1005-1013

[11] Jordan M.I., Jacobs R.A. (1994) Hierarchical mixtures of experts and the EM algorithm, *Neural Comp.*, **6**, 181-214

[12] Keribin C. (2000) Consistent estimation of the order of mixture models, *Sankhya: The Indian Journal of Statistics*, **62**, 49-66

[13] Liu X., Shao Y. (2003) Asymptotics for likelihood ratio tests under loss of identifiability, *The Annals of Statistics*, **31(3)**, 807-832

[14] Olteanu M., Rynkiewicz J. (2008) Estimating the number of components in a mixture of multilayer perceptrons, *Neurocomputing / EEG Neurocomputing*, **71**, 1321-1329

[15] Van der Vaart A.W. (2000) *Asymptotic Statistics*, Cambridge University Press